THE BELL SYSTEM TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

Volume 58 March 1979 Number 3

Approximations to Stochastic Service Systems, with an Application to a Retrial Model

By A. A. FREDERICKS and G. A. REISNER

(Manuscript received February 1, 1978)

This paper illustrates the usefulness of state-dependent, birth-death processes in reducing the dimensions of stochastic service systems. The approximation techniques introduced have wide applicability to general (finite) multidimensional, state-dependent, birth-death processes. These techniques are introduced by considering the "classical" telephony problems dealing with trunk group overflow traffic from the point of view of state-dependent, birth-death processes. The main part of the paper then applies these techniques to a two-dimensional trunk group retrial model of Wilkinson and Radnik. The method, which reduces the W-R model to an approximate, easily-solved, one-dimensional model, makes use of the transition probabilities for state-dependent, birth-death processes. These are obtained via a simple extension of known results. We use the one-dimensional results to compute blocking for a range of parameter values (trunk group sizes and retrial rates) exceeding the computational limits of the W-R model. Maximum relative errors do not exceed 10 to 15 percent, while for most cases of practical interest the relative errors are less than 5 percent. The approximation also provides insight into the region of applicability of even simpler retrial models. This one-dimensional retrial model actually applies to more general (finite) state-dependent, birth-death processes (e.g., loss-delay systems).

I. INTRODUCTION

The purpose of this paper is to illustrate the usefulness of state-dependent birth and death processes in reducing the dimensions of

stochastic systems. The principal application is the reduction of a two-dimensional retrial model, proposed by R. I. Wilkinson and R. C. Radnik,[1] to an approximate one-dimensional model, which is then readily solved. While algorithms[2] exist for the numerical solution of the two-dimensional Wilkinson-Radnik model, the large number of states that are often needed can result in convergence difficulties.

Section II presents the history of and motivation for the techniques used throughout the paper. The equations for the reduced one-dimensional retrial model are described in Section III. Their solution is discussed in Section IV. Section V contains numerical results and comparisons, and Section VI discusses the theoretical accuracy of our one-dimensional approximation.

The main points of this paper are the dimension reduction of stochastic models using a state-dependent birth-and-death process and a method of solution of the resulting approximate model. However, the retrial problem considered here as an example is of interest in itself and has been extensively studied in the past. L. Kosten[3] and J. Riordan[4] considered retrials coming back in a secondary, uncorrelated Poisson stream. J. W. Cohen[5] allowed for negative exponential distributions in the interarrival times of calls, the holding times of calls, the duration of the time interval between two successive attempts by a subscriber whose call was blocked at the first attempt, and the time during which a subscriber continues to make repeated attempts. All these works* attest to the difficulty of modeling and obtaining numerical solutions to the retrial problem. We hope that the ease with which one can obtain reasonable numerical results by using state-dependent, birth-death processes will motivate readers to consider this as one possible approach to the simplification of probabilistic systems.

II. HISTORY AND MOTIVATION FOR THE USE OF STATE-DEPENDENT BIRTH RATES

One early use of state-dependent birth rates to reduce a multidimensional system to a one-dimensional model is given in Ref. 6. The motivating problem was the analysis of an alternate-routed telephone network. The system to be analyzed consists of one or more primary groups of servers, each with its own arrival process. Arrivals which find all servers in its primary group busy are offered to a common overflow group (see Fig. 1). With the common assumption of Poisson traffic for the underlying arrival processes and exponential service times, one can readily write down the appropriate birth-death equations. For the case where there is only one primary group, the analysis

* Those noted are only meant to be representative of various works concerned with the retry phenomenon. They are not meant to provide a complete list of such works.

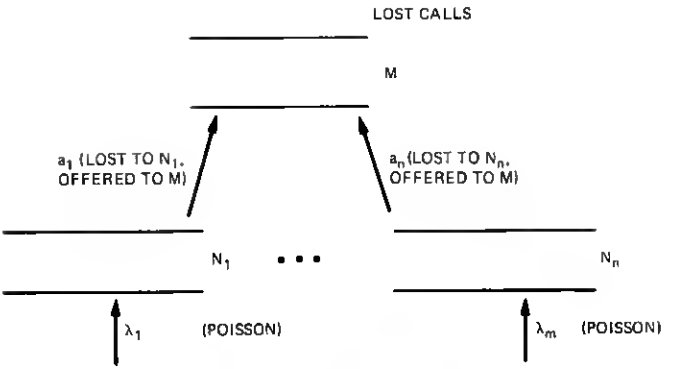


Fig. 1—Overflow problem.

is already somewhat complicated, although it has been carried out.[7] Moreover, reasonably sized trunk groups quickly lead to systems where even numerical solutions are not feasible. However, one is often only interested in the behavior of the overflow group, e.g., finding which attempts are blocked there and hence lost from the combined system. In this case, the primary trunk groups are of interest only inasmuch as they supply the input to the overflow group. It is here that state-dependent, birth-rate modeling has proved of value.* But before noting some of the previous work on state-dependent birth rates related to this overflow problem, it is useful to consider the application of our basic ideas on dimensionality reduction.

For the simplest case of one primary group of N servers overflowing to a group of M servers, Poisson input rate λ, and (unit) exponential holding time, the birth-death equations can be written as

$$\begin{aligned}
(\lambda + i + j)P_{ij} &= \lambda P_{i-1,j} + (j+1)P_{i,j+1} + (i+1)P_{i+1,j} \\
&\qquad i < N \\
(\lambda + N + j)P_{Nj} &= \lambda P_{N-1,j} + (j+1)P_{N,j+1} + \lambda P_{N,j-1} \qquad (1) \\
&\qquad i = N \qquad j < M \\
(N + M)P_{NM} &= \lambda P_{N-1,M} + \lambda P_{N,M-1} \\
&\qquad (P_{ij} = 0 \quad \text{if} \quad i \quad \text{or} \quad j < 0),
\end{aligned}$$

where P_{ij} is the probability that there are i busy servers in the primary group and j busy servers in the secondary group.

Since our main interest is in the marginal distribution $P_{\cdot j}$, we begin

* Actually, if total lost calls were the only item of interest, then the equivalent random method (Ref. 8) would perhaps be more applicable. However, we will be considering more general questions here.

by summing eq. (1) over i. The result after simplification is

$$\begin{aligned} jP_j + \lambda P_{Nj} &= (j+1)P_{j+1} + \lambda P_{N,j-1} \qquad j < M \\ MP_M &= \lambda P_{N,M-1} \qquad j = M, \\ (P_j, P_{N,j} &= 0 \quad \text{if} \quad j < 0), \end{aligned} \tag{2}$$

where we have denoted the marginal distribution $P_{\cdot j}$ by P_j.

If we subtract the equation for $j = M$ from that for $j = M - 1$, and then proceed to subtract the new equation for each j from the old one for $j - 1$ we obtain the equivalent, but simpler, system

$$\lambda P_{Nj} = (j+1)P_{j+1}, \qquad 0 \le j < M. \tag{3}$$

Note that we could have obtained eq. (3) directly by balancing the upward transitions from j to $j + 1$ with the downward transitions from $j + 1$ to j, in equilibrium. In any event, by now using the fact that P_{Nj} can be written as $P_{N|j}P_j^*$ and denoting the term $\lambda P_{N|j}$ by λ_j, we obtain

$$\lambda_j P_j = (j+1)P_{j+1}\ 0 \le j < M, \tag{4}$$

an apparent one-dimensional birth-death process. Care must be taken in this interpretation. The quantity λ_j is the average "birth" rate when there are j busy on the overflow group. That is, the input process can be characterized by a state-dependent birth rate only in an average sense. Thus, while (4) is a valid equation satisfied by the equilibrium probabilities P_j, other quantities that might be obtained by viewing this as a birth-death process (e.g., transitory behavior) would at best be approximate. Indeed, even to obtain the P_j's from (4) it would be necessary to determine the λ_j exactly. Since this can usually only be done by solving the combined system (primary plus overflow trunk group), a problem we wish to avoid, we turn to approximations for the λ_j's.

Linear birth rates (e.g., $\lambda_j = a + bj$) were suggested in Ref. 6 for the case where the overflow group is infinite. This results in a negative binomial distribution for the state probabilities. Determining the parameters a and b by matching the mean and variance of the number of busy servers was found to result in a reasonably good approximation for the state probabilities. The idea of relating the birth rates to conditional probabilities was presented in Ref. 7. The resulting approximation for the state probabilities for an infinite overflow group were found to be better than those obtained with the negative binomial distribution, particularly for large values of the number busy.

Extensions of the negative binomial approximation (linear birth

* The notation $P_{N|j}$ refers to the conditional probability of N (busy on the primary group) given j (busy on the overflow group).

rates) to a system with a finite overflow group are given in Refs. 8, 9, and 10. The approach in Ref. 9 is to solve for the equilibrium probabilities $P_i^{(\infty)}$ for an infinite overflow group, and then simply terminate and normalize these to approximate the state probabilities $P_i^{(N)}$ for the finite group, i.e.,

$$P_i^{(N)} = \frac{P_i^{(\infty)}}{\sum_{i=0}^{N} P_i^{(\infty)}}. \tag{5}$$

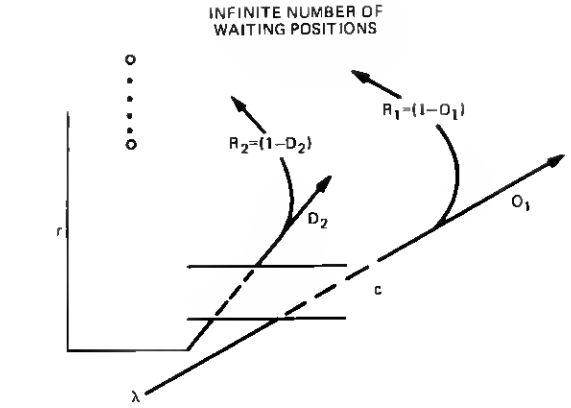
Equation (5) still implies a linear birth rate of the form $\lambda_i = a + bi$. The problem is that if the parameters a, b are adjusted to match moments on an infinite trunk group, then for the finite case the total offered load $\Lambda = \sum_{i=0}^{N} \lambda_i P_i^{(N)}$ will no longer match the true offered load. This problem can be rectified by adjusting the offered load (see, for example, Refs. 9 and 10); however, this generally results in the need for an iteration procedure. For example, a choice of λ_i results in a set of probabilities $P_i^{(N)}$, and hence an actual offered load $\Lambda = \sum_{i=0}^{N} \lambda_i P_i^{(N)}$, which, if not the desired value, results in a need to adjust the λ_i. This feedback effect, where the equilibrium probabilities must be used to adjust the offered load, is a dominant feature in this type of approach to dimensionality reduction. We will see shortly that this interaction between the reduced state probabilities and the assumed state-dependent offered load is even stronger for the retrial model considered.

An important point to note is that, independent of the initial motivation for the above approximations, they can all be interpreted as attempting to approximate the conditional probability that the primary group is busy, given that there are j busy on the secondary group. This interpretation is important since it can often lead to insight into the applicability of the resulting approximation.

Before proceeding we note that state-dependent birth rates have also been used to study multilink systems offered overflow traffic and to obtain approximations for the blocking seen by the various parcels of traffic offered to an overflow group as depicted in Fig. 1.[11-13]

III. APPLICATION TO THE WILKINSON-RADNIK RETRIAL MODEL

A diagram of the Wilkinson-Radnik (W-R) retrial model is given in Fig. 2. The underlying offered load is assumed to be Poisson with rate λ. If these attempts find all c servers busy, they may defect from the system (probability D_1) or they may wait a period of time and then retry (probability $R_1 = 1 - D_1$). Thus, the number of people j waiting to retry is increased by one with probability R_1, whenever a first offered attempt arrives to find the number of busy servers i equal to c. When a customer retries, he either finds an idle server ($i < c$) and



λ – POISSON FIRST OFFERED LOAD
c – NUMBER OF SERVERS
O_1 – PROBABILITY OF DEFECTION FOR FIRST BLOCKED ATTEMPT
O_2 – PROBABILITY OF DEFECTION FOR SUBSEQUENT BLOCKED ATTEMPT
r – RETRIAL RATE (EXPONENTIAL DISTRIBUTION)

Fig. 2—Wilkinson-Radnik retrial model.

hence is carried by the system, or, if $i = c$, he may defect with probability D_2 or again wait and retry with probability $R_2 = 1 - D_2$. With the assumption of exponential times to retry, this system is completely characterized by the state (i,j), i = number of busy servers ($i = 0, \cdots, c$), j = number of customers waiting to retry ($j = 0, \cdots, \infty$).

Denoting the mean time to retry by $1/r$ and the mean server holding time by $1/\mu$, we can readily write the state equations for the probabilities $P_{ij} = P^*$ (i busy server, j waiting to retry). Assuming for simplicity that $R_1 = R_2 = R$, we obtain:

$$\begin{aligned}(\lambda + rj + \mu i)P_{ij} &= \lambda P_{i-1,j} + r(j+1)P_{i-1,j+1} \\ &\quad + \mu(i+1)P_{i+1,j} \qquad i < c \\ (\lambda R + rj(1-R) + \mu c)P_{cj} &= \lambda P_{c-1,j} + r(j+1)P_{c-1,j+1} \\ &\quad + r(j+1)(1-R)P_{j+1} + \lambda R P_{c,j-1} \\ &\quad (P_{i,j} = 0 \quad \text{if} \quad i \quad \text{or} \quad j < 0).\end{aligned} \tag{6}$$

While the W-R model is a reasonable one for the customer retry phenomenon in telephone systems, eq. (6) quickly lead to numerical problems, even for modest values of c. Here, as with the overflow problem, we see that the dimensionality difficulty is caused by an

* P denotes probability.

aspect of the vector state (i, j), namely j, that we often are not concerned with, except as it influences i, the number of busy servers. Hence, following the procedure outlined in Section II, we sum (6) with respect to j and obtain

$$(\lambda + rE(j|i) + \mu i)P_i = \lambda P_{i-1} + rE(j|i-1)P_{i-1} + \mu(i+1)P_{i+1} \qquad i \neq c \quad (7)$$

and

$$(\mu c)P_c = \lambda P_{c-1} + rE(j|c-1)P_{c-1},$$

where

$$P_i = \sum_j P_{ij}$$

$$E(j|i) = jP_{ij}/P_i.$$

The term $\lambda_i' = rE(j|i)$ represents the mean input intensity associated with the retries. If this quantity were known exactly, then (7) could be solved to yield the exact solution for the equilibrium state probabilities P_i. The use of (7) to compute other quantities such as transitory probabilities would again be an approximation. However, one would expect such an approximation to be good if in the two-dimensional model, the value of j (number of retry sources) did not vary much from its mean for a given value of i (number of servers busy). This idea will be explored later.

Before discussing how to obtain an approximation for the λ_i' which clearly depends on the unknown $P(i)$, we note that direct balance of flows across the $(i, i+1)$-state boundary as before yields the simpler (but equivalent to (7)) state equations

$$\lambda_i P_i = \mu(i+1)P_{i+1}, \qquad (8)$$

where

$$\lambda_i = \lambda + \lambda_i' = \lambda + rE(j|i).$$

It is clear that there is a strong interaction between the state probabilities P_i and the retry intensity λ_i'. We now turn our attention to modeling this rather complicated relationship, and hence obtaining a solution to (8) which will hopefully approximate the two-dimensional W-R retry model adequately.

IV. SOLUTION OF THE ONE-DIMENSIONAL RETRIAL MODEL

The "solution" to a one-dimensional birth-death process is, of course, well known, provided the birth (and death) rates are *given*. Thus the main problem we are faced with is determining the λ_i's for the one-dimensional model so that they capture the essence of the two-dimen-

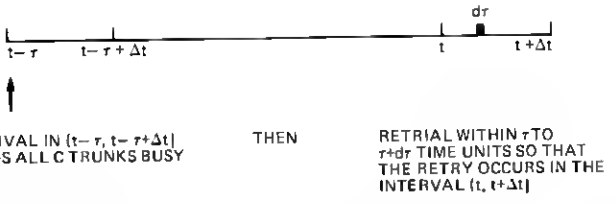


Fig. 3—Necessary events for a retrial arrival in the interval $(t, t + \Delta t)$ conditioned on an inter-retrial time equal to τ.

sional model. As indicated earlier, λ_i is the sum of λ (first offered traffic intensity) and λ_i' (retrial intensity when i trunks are busy). We thus have, to first-order in Δt, that the λ_i''s satisfy the relationship

$$\lambda_i' \Delta t = P^{(1)}\{\text{exactly one retrial in } (t, t + \Delta t] \mid N(t) = i\}, \quad (9)$$

where $N(t)$ is the number of busy servers at time t and the superscript (1) indicates that this probability is for the one-dimensional model.

What we would like is for $P^{(1)}$ in (9) to be close (in some sense) to $P^{(2)}$, the corresponding probability for the two-dimensional model. Using the law of total probability, conditioned on the inter-retrial time of the arrival under consideration, we have*

$$P^{(2)}\{\text{exactly one retrial in } (t, t + \Delta t] \mid N(t) = i\}$$

$$= \int_\tau P\{N(t - \tau) = c, a(t - \tau, \Delta t), r(\tau, d\tau) \mid N(t) = i\}, \quad (10)$$

where $a(t - \tau, \Delta t)$ is the event that an arrival occurs in $(t - \tau, t - \tau + \Delta t]$ and $r(\tau, d\tau)$ is the event that he will retry if blocked within $(\tau, \tau + d\tau]$ time units. Figure 3 represents these events pictorially. (Note that the assumptions made for the two-dimensional model imply that each blocked arrival can be tagged with a time to retrial, τ, taken independently from a distribution $F_r(\tau)$, at the time of his arrival.)

Using $P(A \mid B) = P(B \mid A)P(A)/P(B)$ in (10) we obtain

$$P^{(2)}\{\cdot\} = \int_\tau P\{N(t) = i \mid N(t - \tau) = c, a(t - \tau, \Delta t), r(\tau, d\tau)\}$$

$$\cdot P\{N(t - \tau) = c, a(t - \tau, \Delta t), r(\tau, d\tau)\} / P\{N(t) = i\}. \quad (11)$$

Denoting the events $\{a(t - \tau, \Delta t), r(\tau, d\tau)\}$ by $\{A_r\}$ and using the law of total probability conditioned on $J(t - \tau)$ and $J(t)$, the number waiting to retry at times $t - \tau$ and t, result in

* For simplicity, we omit all terms of higher order than Δt in eqs. (10) to (16).

$$P^{(2)}\{\cdot\} = \int_\tau \sum_{j_1=0}^{\infty} \sum_{j_2=1}^{\infty}$$

$$\cdot P\{N(t) = i, J(t-\tau) = j_1, J(t) = j_2 \mid N(t-\tau) = c, A_r\}$$

$$\cdot P\{A_r\}/p\{N(t) = i\}. \quad (12)$$

Using $P(A, B \mid C) = P(A \mid B, C)P(B \mid C)$, this becomes

$$P^{(2)}\{\cdot\} = \int_\tau \sum_{j_1=0}^{\infty} \sum_{j_2=1}^{\infty} \cdot P\{N(t) = i, J(t)$$

$$= j_2 \mid N(t-\tau) = c, J(t-\tau) = j_1, A_r\}$$

$$\cdot \frac{P\{J(t-\tau) = j_1 \mid N(t-\tau) = c, A_r\} P\{A_r\}}{P\{N(t) = i\}}. \quad (13)$$

Now the value of $J(t-\tau)$ represents the retrial intensity at $(t-\tau)$. In general, for the two-dimensional model, $P\{J(t-\tau) = j_1\}$ does indeed depend not only on the value of $N(t-\tau)$, but on the fact that an arrival has just occurred. However, this latter dependence is inconsistent with the one-dimensional model. More specifically, we have assumed that the retrial intensity depends only on the state of the one-dimensional system, $N(t-\tau)$. Thus we are led to making the approximation

$$P\{J(t-\tau) = j_1 \mid N(t-\tau) = c, A_r\}$$

$$= P\{J(t-\tau) = j_1, N(t-\tau) = c\}.$$

Note that this approximation should tend to underestimate the retrial intensity λ_c' when there are c busy servers and hence underestimate the blocking, particularly as seen by the retrials.

Using this approximation, the formula $P(A \mid B, C)P(B \mid C) = P(AB \mid C)$, and noting that

$$p\{N(t) = i, J(t) = j_2 \mid N(t-\tau) = c, J(t-\tau) = j_1, A_r\}$$

$$= P\{N(t) = i, J(t) = j_2 - 1 \mid N(t-\tau) = c, J(t-\tau) = j_1\},$$

we obtain

$$P^{(2)}\{\cdot\} \approx \int_\tau \sum_{j_1=0}^{\infty} \sum_{j_2=0}^{\infty} P\{N(t) = 1, J(t) = j_2, J(t-\tau)$$

$$= j_1 \mid N(t-\tau) = c\}$$

$$\cdot P\{N(t-\tau)=c\}P\{A_r\}/P\{N(t)=i\}$$

$$=\int_\tau P\{N(t)=i\,|\,N(t-\tau)=c\}\,P\{N(t-\tau)=c\}$$

$$\cdot P\{A_r\}/P\{N)t)=i\}. \tag{14}$$

Thus we require that

$$\lambda_i'\Delta t=\int_\tau P^{(1)}\{N(t)=i\,|\,N(t-\tau)$$

$$=c\}\,P^{(1)}\{N(t-\tau)=c\}\,P^{(1)}\{A_r\}/P^{(1)}\{N(t)=i\}, \tag{15}$$

where we have used the superscript (1) to emphasize that the probabilities are for the one-dimensional model. Using the Markovian properties and independence assumptions for a birth-death process, (15) can be written as

$$\lambda_i'\Delta t=\int_\tau P_{ci}(\tau)\;(\lambda_c\Delta t)\,P_cRdF_r(\tau)P_i, \tag{16}$$

where $P_{ci}(\tau)=P\{N(\tau)=i\,|\,N(\mathrm{o})=c\}$, i.e., the transition probabilities

λ_c = birth rate when $N=c$

R = Pr {a blocked attempt will retry}

$F_r(\tau)$ = distribution function for the time of retry

$P_c=P\{N=c\}$
$P_i=P\{N=i\}$ equilibrium probabilities.

All are from the one-dimensional model.

Thus, finally, we have that the overall state dependent birth rates are given by

$$\lambda_i=\lambda+\lambda_i', \tag{17}$$

where from (16)

$$\lambda_i'=\frac{P_c\lambda_cR}{P_i}\int_\tau P_{c_i}(\tau)\;dF_r(\tau). \tag{18}$$

Up to this point, we have not made any assumptions regarding the retrial distribution F_r. Before doing so, it is worth pointing out that eq. (18) provides the correct answer for the two limiting values of the retrial rate r. As r goes to infinity, retrials occur "immediately." If we represent this by setting F_r equal to a unit step at zero, then (18) reduces to

$$\lambda_i'=\frac{P_c\lambda_cR}{P_i}\,\delta_{ci}, \tag{19}$$

where δ_{ci} is the Kronecker delta function. Combining eqs. (17) and (18) we find that $\lambda_i = \lambda$ for i less than c, and $\lambda_c = \lambda/(1 - R)$. Thus, when retrials occur immediately, they do not influence the state of the system. With one or more servers idle, the load remains at λ; with all servers occupied, an arrival retries infinitely fast and (for $R < 1$) will exit from the system before a server becomes idle.

In the other limiting case, as r goes to zero, customers "come back in an uncorrelated stream." Proceeding as above, and noting that

$$\int_0^\infty P_{ci}(\tau)\, dF_r(\tau) = P_i \qquad \text{as } r \to 0$$

(if we let F_r approach a unit jump function at infinity), we find that $\lambda_i = \lambda + P_c\lambda_c R$ for all i, and thus $\lambda_i = \lambda/(1 - P_c R)$. That is, the traffic intensity increases by a factor of $(1 - P_c R)^{-1}$, independent of the number of servers occupied. This is a familiar retrial model. In Section V we see that there is a wide range of parameter values for which this simple model is not a useful approximation for the proportion of retrials blocked.

We now return to eqs. (17) and (18), and use them to calculate recursive formulas for the λ_i's. We assume, as in the two-dimensional retrial model, that the time to retrial is given by a negative exponential distribution with rate r, namely

$$F_r(\tau) = 1 - e^{-r\tau}. \tag{20}$$

In this case, the equations become

$$\lambda_i = \lambda + \frac{P_c\lambda_c R}{P_i}\, r\hat{P}_{ci}(r) \qquad i = 1, \cdots, c, \tag{21}$$

where $\hat{}$ denotes the Laplace transform. In particular,

$$\lambda_c = \frac{\lambda}{1 - rR\hat{P}_{cc}(r)}. \tag{22}$$

If we let

$$\delta_j = \frac{\lambda_j - \lambda}{\lambda_{j-1} - \lambda} \qquad j = 1, \cdots, c, \tag{23}$$

then we can write

$$\lambda_{j-1} = \lambda + \frac{\lambda_j - \lambda}{\delta_j} \qquad j = 1, \cdots, c. \tag{24}$$

Substituting (17) into (19) and simplifying using $\lambda_{j-1}P_{j-1} = \mu_j P_j$, we obtain

$$\delta_j = \frac{\mu_j \hat{P}_{cj}(r)}{\lambda_{j-1}\hat{P}_{cj-1}(r)} \qquad j = 1, \cdots, c. \tag{25}$$

Using formula for $\hat{P}_{cj}(r)$ developed in the appendix,* we obtain a recursion for δ_j:

$$\delta_1 = 1 + r/\lambda_0 \quad \text{and} \quad \delta_j = \frac{(r + \lambda_{j-1} + \mu_{j-1})\delta_{j-1} - \mu_{j-1}}{\lambda_{j-1}\delta_{j-1}} \quad \text{for} \quad j = 2, \cdots, c. \tag{26}$$

Combining (26) with (22), (24), and the formula for $\hat{P}_{cc}(r)$,

$$\hat{P}_{cc}(r) = \frac{\delta_c}{(r + \mu_c)\delta_c - \mu_c}, \tag{27}$$

we get an iteration scheme for finding the $\lambda_i i = 0, \cdots, c$ which satisfies the birth-death equations and is consistent with the derivation of the $S_{jj} = 0, \cdots, c$.

Before leaving this section, we note that the above derivation holds equally well if the underlying system (without retries) is characterized by an arbitrary (finite) state-dependent, birth-death process. For example, it applies to the loss delay system considered in Ref. 4, and to overflow traffic characterized via state-dependent birth rates (as discussed in Section II). Moreover, using the results given in the appendix, one can compute transition probabilities and other related quantities for this system (e.g., correlation function).

V. NUMERICAL RESULTS

We assess the accuracy of the one-dimensional approximation by comparing our results to those obtained from direct numerical solution of the two-dimensional W-R model. The particular comparisons presented here were chosen to give the reader an understanding of the value of the one-dimensional approximation for a wide range of parameter values. However, due to the difficulty of computing "correct" values (i.e., from the two-dimensional model), the results may not cover every possible region of interest. In particular, it is difficult to analyze the convergence behavior of the two-dimensional algorithm for large trunk groups or small retrial rates. This difficulty arises because the number of waiting positions must increase to obtain a good approximation to an infinite waiting room; in turn, the number of states increases markedly and roundoff errors may become significant. Fortunately (see Section VI), the two models give the same results as r tends to infinity or zero. The approximation is worst for values of r around 2 and gets progressively better as r gets larger or smaller.

* The appendix shows that the $\hat{P}_{ij}(t)$ of a general one-dimensional birth-death process may be obtained in precisely the same way they were obtained for a combined delay and loss system in Ref. 4. This fact was recognized and used in Ref. 14 (Appendix B).

We first look at retrial blocking, i.e., the proportion of reattempts blocked. Figure 4 shows the retrial blocking for a system with two servers, as a function of offered load. We assume a probability of $R = 0.8$ that a blocked customer will retry. As expected, the proportion of reattempts blocked increases as r, the retrial rate, increases. For the two cases shown, r equal to 2.0 and 0.5, the relative difference between the two models is approximately 10 to 15 percent (as noted earlier, this is the worst case). We also see that the retrial blockings for the two retrial rates approach one another as offered load increases.

For comparison, we show similar retrial blocking curves for systems with 5 and 30 servers, in Figs. 5 and 6, respectively. An interesting phenomenon can be seen here, but first observe that the offered loads in each of Figs. 4 through 6 correspond to values of call congestion ranging from 0.01 to 0.30, without considering retrials. For a given design load, without retrials, the percentage of reattempts blocked is much higher for a smaller number of servers. In particular, at 1-percent total blocking and $r = 2.0$, the retrial blocking is 51 percent for 2 servers and 17 percent for 30 servers. This poor retrial performance of small server groups may be of importance in understanding customer satisfaction (or annoyance). We make one last point regarding retrial

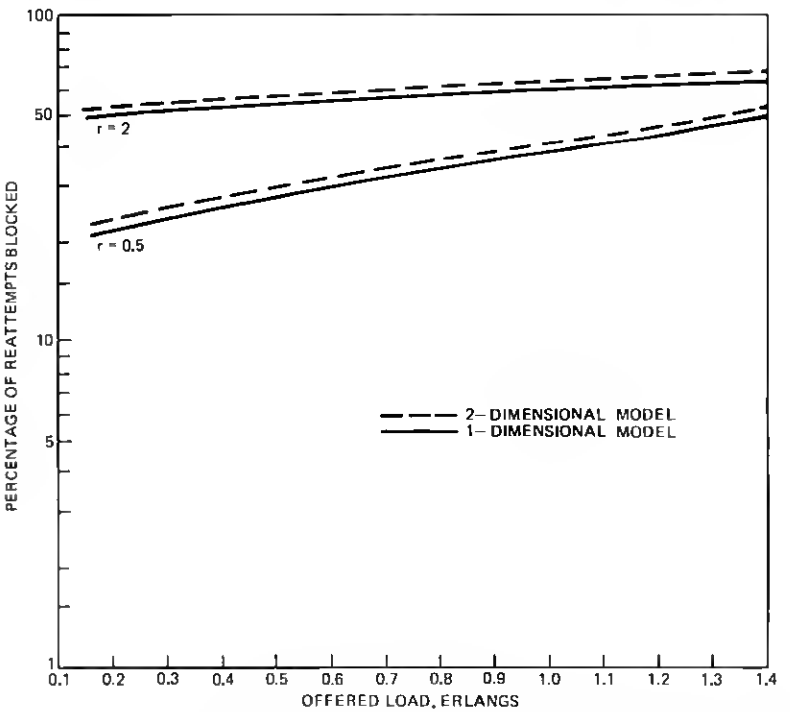


Fig. 4—Retrial blocking, 2 servers, retrial probability = 0.8.

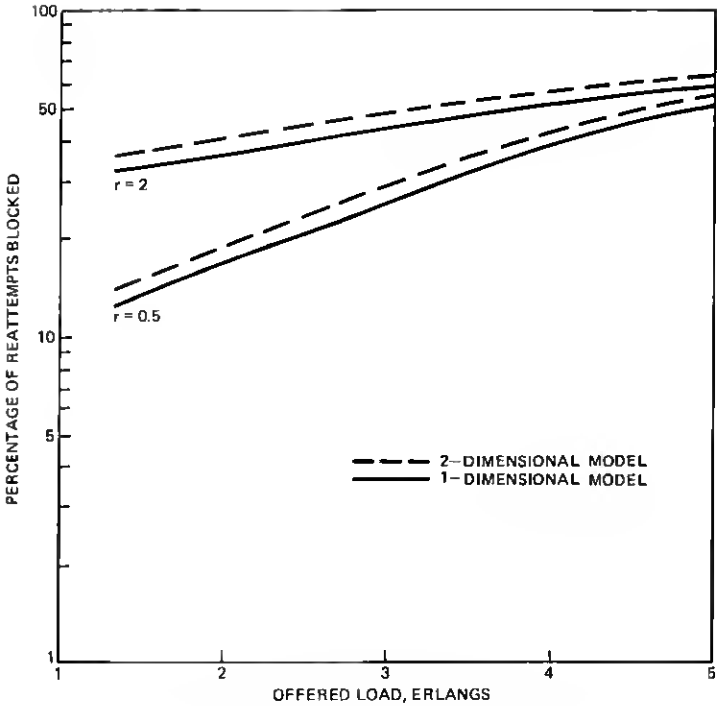


Fig. 5—Retrial blocking, 5 servers, retrial probability = 0.8.

blocking by referring the reader back to Fig. 6. The curve for $r = 0$ corresponds to the simply-computed "uncorrelated retrial" model. We see that, for a high-design blocking, both the one- and two-dimensional models are well approximated by this simpler model.

Figure 7 illustrates the dependence of retrial blocking and call congestion on trunk group size. The proportion of retrials blocked generally decreases as the group size N increases and seems to approach an asymptote. The total call congestion also decreases for small to medium size groups, but then increases for N larger than 30 and approaches the same asymptote. Notice that, if we restricted our attention to the region where the two-dimensional model applies, we would not get a good view of the limiting behavior. The reason for the seemingly anomalous behavior of total call congestion is that the larger and exceedingly efficient trunk groups are correspondingly more sensitive to traffic above the design load. We now briefly discuss the limiting behavior.

If we assume that retrials return in an uncorrelated Poisson stream ($r = 0$ case), then the inflated load Λ is given by the solution (determined by iteration, e.g.,) to

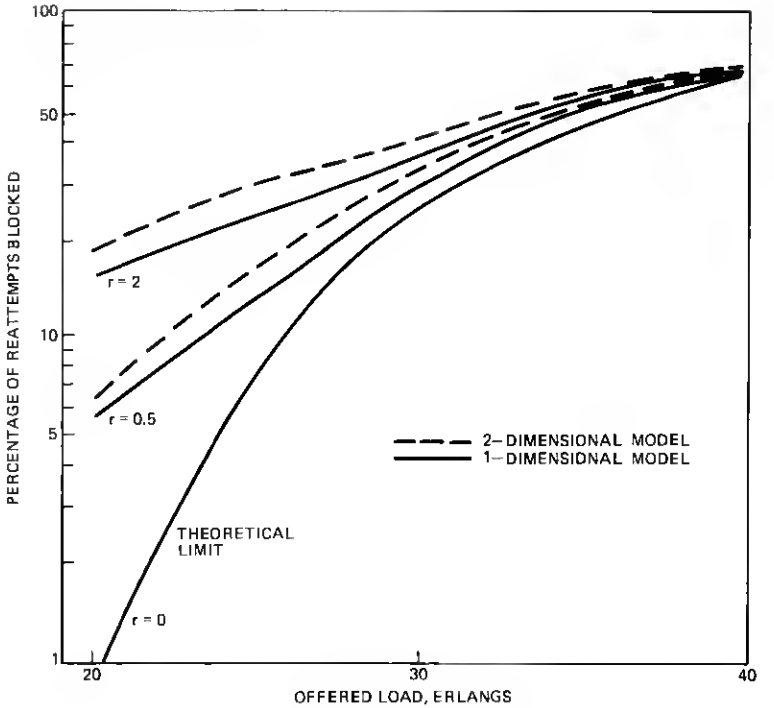


Fig. 6 —Retrial blocking, 30 servers, retrial probability = 0.8.

$$\Lambda = \frac{\lambda}{1 - B(N, \Lambda) \cdot R}, \tag{28}$$

where $B(N, \Lambda)$ is the Erlang-B blocking function. The resulting blocking (for both retrials and first attempts) is the lowest curve in Fig. 7. For large values of N, we use the well-known approximation $1 - N/\lambda$ for $B(N, \lambda)$ (see Ref. 15) in conjunction with (28) to obtain the asymptote,

$$\text{call congestion} = \frac{B}{1 - R(1 - B)}, \tag{29}$$

for a fixed design blocking B. In Fig. 7, the asymptote 0.048 is shown by a dashed line.

Another way of obtaining (29) is to assume that, for large trunk groups, $\lambda - N$ equals the traffic lost on first attempts, $(\lambda - N)R$ equals the traffic lost on second attempts, $\cdots$, $(\lambda - N)R^{k-1}$ equals the traffic lost on the kth attempts, etc. Then the total blocking is given by

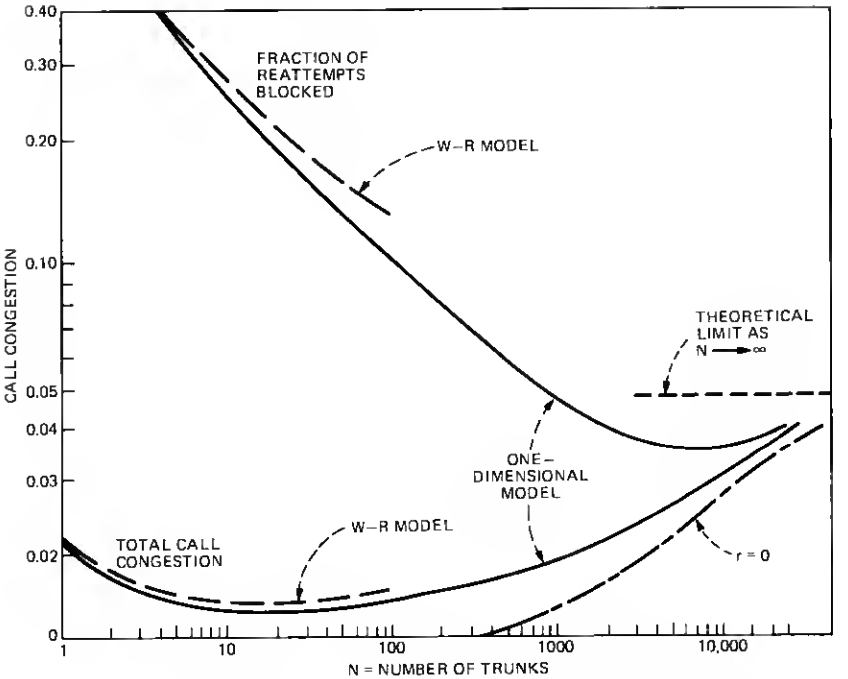


Fig. 7—Call congestion with retrials for trunk groups engineered at B.01.

$$\frac{\text{lost}}{\text{offered}} = \frac{(\lambda - N) + (\lambda - N)\left(\dfrac{R}{1-R}\right)}{\lambda + (\lambda - N)\left(\dfrac{R}{1-R}\right)} = \frac{1 - N/\lambda}{1 - R(N/\lambda)} = \frac{B}{1 - R(1-B)}.$$

We close this section by pointing out that the one- and two-dimensional models yield practically the same values for the call congestion and the time congestion. A few sample values are shown in Table I.

VI. THEORETICAL ACCURACY—A NUMERICAL LOOK

As indicated in the derivation in Section III, the one-dimensional state equations may be obtained by summing the two-dimensional state equations over the number of waiting positions occupied. When this is done, the arrival rate of retries (when i servers are busy) equals the retrial rate times the expected number of waiting positions occupied (given i servers busy). In equation form,

$$\lambda_i' = r \cdot E(j \mid i). \tag{30}$$

If one knew the conditional expectations $E(j \mid i)$ precisely, then all the λ_i's, and hence all the P_i's of the one-dimensional equations, would be known exactly. The approximation occurs in the iteration procedure for finding the λ_i's where we implicitly assume that the one-dimensional equations actually come from a one-dimensional Markov model that adequately describes the retrial situation.

Intuitively, if either (i) the standard deviation-to-mean ratio $\sigma_{j \mid i}/E(j \mid i)$ is very small, or if (ii) $E(j \mid i) \pm k\sigma_{j \mid i}$ is "equivalent" to $E(j \mid i)$ (for some reasonable k), then for practical purposes we have a proper one-dimensional system; since knowing the number of busy servers i implies that we know j (namely, $E(j \mid i)$). Hence, if either of conditions (i) or (ii) above hold, then we expect the one- and two-dimensional models to yield similar results. We look at a numerical example to illustrate the point. Table II shows the effect of varying the retrial rate, for five servers offered 2.22 Erlangs, while keeping all other parameters fixed. As the retrial rate goes to zero, the ratio $\sigma_{j \mid 5}/E(j \mid 5)$

Table I—Time and call congestion (r = 2.0, R = 0.8)

		One-Dimensional Model		Two-Dimensional Model	
Offered Load	No. of Servers	Time Cong	Call Cong	Time Cong	Call Cong
0.381	2	0.055	0.089	0.055	0.092
2.22	5	0.062	0.084	0.063	0.088
10.6	15	0.072	0.085	0.073	0.087
24.8	30	0.081	0.092	0.084	0.100
0.595	2	0.116	0.179	0.116	0.186
2.88	5	0.136	0.176	0.136	0.188
12.5	15	0.169	0.195	0.167	0.204
28.1	30	0.194	0.212	0.197	0.227

Table II—Effect of retrial rate (c = 5 servers, offered load = 2.22 Erlangs)

Retrial Rate	$E(j \mid 5)$	λ'_5/r	$\sigma_j \mid 5$	$\sigma_j \mid 5/E(j \mid 5)$
128.0	0.058	0.058 (0%)*	0.24	4.14
64.0	0.102	0.101 (1%)	0.32	3.14
32.0	0.166	0.161 (3%)	0.42	2.53
16.0	0.250	0.234 (6%)	0.52	2.08
8.0	0.349	0.316 (9%)	0.64	1.83
4.0	0.460	0.403 (12%)	0.76	1.65
2.0	0.577	0.499 (14%)	0.88	1.53
1.0	0.702	0.609 (13%)	1.00	1.42
0.5	0.856	0.758 (11%)	1.14	1.33
0.25	1.095	0.997 (9%)	1.31	1.20
0.125	1.527	1.440 (6%)	1.55	1.02
0.0625	2.398	2.300 (4%)	1.99	0.83
0.03125	4.107	4.010 (2%)	2.62	0.64
0.015625	7.515	7.422 (1%)	3.55	0.47

* The number in parentheses is the relative difference between $E(j \mid 5)$ and our one-dimensional approximation to it, λ'_5/r.

tends to zero [as do $\sigma_{j|i}/E(j|i)$, $i = 0, 1, \cdots, 4$—not shown in Table II]. At the same time, our approximation λ_5'/r converges to $E(j|5)$.

At the other extreme, as the retrial rate tends to infinity, Table II shows again that λ_5'/r converges to $E(j|5)$. However, the standard deviation-to-mean ratio tends to infinity, proving that this is not a sufficient condition for convergence. On the other hand, $E(j|5) + k\sigma_{j|5} \simeq E(j|5)$, for any k. Intuitively, if we ignore the state $i = c$ $(= 5)$, we again have a one-dimensional Markovian system. Indeed, as r increases retrials occur instantaneously and the probability of having anyone in the waiting room tends to zero, if any server is free.

In summary, λ_c/r converges to $E(j|c)$ as r tends to zero or infinity. In these cases, the one-dimensional model gives the exact same answer as the W-R model but is much easier to compute (especially for r small since in this case a very large waiting room is needed). Unfortunately, the approximation seems worst for values of r around 2.0. Nevertheless, Table II shows only a 14-percent relative error in the approximation for $E(j|5)$ in this case.

VII. ACKNOWLEDGMENTS

This work has benefited greatly from many helpful discussions with A. E. Eckberg, J. M. Holtzmann, and J. S. Kaufman. We would also like to thank R. H. Harris and H. D. Jacobsen for assistance in obtaining needed data from the two-dimensional W-R retrial model.

APPENDIX

Transition Probabilities of a One-Dimensional Birth-Death Process

We show that one may obtain a solution for the transition probabilities for any finite-state, one-dimensional, birth-death process. The solution is not original; it precisely follows the derivation of the transition probabilities for the simplest combined delay and loss system given by J. Riordan (Ref. 4, pp. 96–98).

Assume we have arbitrary state-dependent birth and death rates, λ_i and μ_i, respectively. Further assume that $\lambda_i = 0$ for $i \geq c + 1$ and define $\mu_0 = \lambda_{-1} = \mu_{c+1} = 0$. Then one obtains the usual system of ordinary differential equations

$$P_{ij}'(t) = \lambda_{j-1}P_{ij-1}(t) - (\lambda_j + \mu_j)P_{ij}(t) + \mu_{j+1}P_{ij+1}(t)$$
$$k, j = 0, 1, \cdots, c. \quad (31)$$

Since $P_{ij}(0) = 0$ for $i \neq j$ and $P_{ii}(0) = 1$, the Laplace transform of $P_{ij}'(t)$ is given by $sP_{ij}(s)$ for $i \neq j$ and by $-1 + sP_{ij}(s)$ for $i = j$. Hence, the Laplace transform of eq. (31) is

$$\lambda_{j-1}\tilde{P}_{ij-1}(s) - (s + \lambda_j + \mu_j)\tilde{P}_{ij}(s) + \mu_{j+1}\tilde{P}_{ij+1}(s) = \delta_{ij}$$
$$i, j = 0, 1, \cdots, c. \quad (32)$$

For any $i = 0, 1, \cdots, c$ we can write

$$D\Pi_i = \delta_i, \tag{33}$$

where δ_i is a $c + 1$ dimensional vector whose $(i + 1)$st component is 1, with all other components equal to zero,

$$\Pi_i = \begin{bmatrix} P_{i0} \\ P_{ij} \\ P_{ic} \end{bmatrix},$$

and

$$D = \begin{bmatrix} s+\lambda_0 & -\mu_1 & 0 & \cdot & \cdot & 0 & 0 \\ -\lambda_0 & s+\lambda_1+\mu_1 & -\mu_2 & 0 & & & 0 \\ 0 & -\lambda_2 & s+\lambda_2+\mu_2 & -\mu_3 & 0 & & \\ \cdot & & & & & & \\ \cdot & & & & & & 0 \\ \cdot & & & 0 & -\lambda_{c-2} & s+\lambda_{c-1}+\mu_{c-1} & -\mu_c \\ 0 & 0 & & & 0 & -\lambda_{c-1} & s+\mu_c \end{bmatrix}.$$

If we define D_i, Δ_i via

$$D_0 = 1$$

$$D_1 = s + \lambda_0$$

$$D_{i+1} = (s + \lambda_i + \mu_i)D_i - \lambda_{i-1}\mu_i D_{i-1} \qquad 1 = 2, \cdots, c$$

and

$$\Delta_0 = 1$$

$$\Delta_1 = s + \mu_c$$

$$\Delta_{i+1} = (s + \lambda_{c-i} + \mu_{c-1})\Delta_i - \lambda_{c-1}\mu_{c-i+1}\Delta_{i-1},$$

then the transforms, $\hat{P}_{ij}(s)$, of the transition probabilities are given by

$$\mathrm{Det}(D)\hat{P}_{ij}(s) = \begin{cases} \lambda_i\lambda_{i+1} \cdots \lambda_{j-1}D_i\Delta_{c-j} & i < j \\ \mu_{j+1}\mu_{j+2} \cdots \mu_i D_j \Delta_{c-i} & i > j, \\ D_i\Delta_{c-i} & i = j \end{cases} \tag{34}$$

where the determinant of D can be written as

$$\mathrm{Det}(D) = D_c(s + \mu_c) - \lambda_{c-1}\mu_{nc}D_{c-1}.$$

Equation (34) is the key formula used in Section IV.

We can go one step further and write down a "closed-form" solution for the $P_{ij}(t)$, namely,

$$P_{ij}(t) = \sum_r \frac{D_i(r)D_j(r)}{\mathcal{C}_i M_i S_c(r)} e^{rt},$$

where

$$\ell_i = \lambda_0 \cdots \lambda_{i-1}$$

$$M_j = \mu_0 \cdots \mu_i$$

$$S_c(r) = \sum_{i=0}^{c} \frac{D_i^2(r)}{\ell_i M_i}$$

and the sum is over the c roots of $\det(D) = 0$. Of course, the use of this solution is limited by one's ability to find characteristic roots.

REFERENCES

1. R. I. Wilkinson and R. C. Radnik, "Customer Retrials in Toll Circuit Operation," Conference Record, 1968 IEEE International Conference on Communications, June 12–14, 1968, Philadelphia, Pa.
2. R. H. Harris and S. R. Neal, unpublished work, and H. D. Jacobsen, unpublished work.
3. L. Kosten, "Over de invloed van herhaald oproepen in de theorie der blokkeringskansen" (On the Influence of Repeated Calls in the Theory of Probabilities of Blocking), De Ingenieur, *47* (1947), pp. 1–25.
4. J. Riordan, *Stochastic Service Systems* (1962).
5. J. W. Cohen, "Basic Problems of Telephone Traffic Theory and the Influence of Repeated Calls," Philips Telecommunications Review, *18*, No. 2 (August 1957).
6. R. I. Wilkinson, "Theories for Toll Traffic Engineering in the U.S.A." B.S.T.J., *35*, No. 2 (March 1956), pp. 421–514.
7. E. Brockmeyer, "The Simple Overflow Problem in the Theory of Telephone Traffic," Teleteknik, *5*, 1954, pp. 361–374.
8. B. Wallstrom, "A Distribution Model for Telephone Traffic with Varying Call Intensity, Including Overflow Traffic," Ericsson Technics, *20*, No. 2 (1964), pp. 183–202.
9. B. Wallstrom, "Cungestion Studies in Telephone Systems with Overflow Facilities," Ericsson Technics, 22, No. 3 (1966).
10. R. R. Mina, "Some Practical Applications of Teletraffic theory," Fifth International Teletraffic Conference, 1967, pp. 428–434 (appendix by R. Syski).
11. J. M. Holtzman, "Point-to-Point Blocking Probabilities for Non-Poisson Traffic," unpublished work.
12. A. A. Fredericks, "On the Determination of Individual Parcel Blocking Probabilities for Overflow Traffic via State Dependent Birth Rates," unpublished work.
13. A. A. Fredericks, "A New Approach to Parcel Blocking via State Dependent Birth Rates," unpublished wurk.
14. A. A. Fredericks, "Impact of Traffic Factors and Other System Parameters on TASI-D Performance," unpublished work.
15. R. Syski, *Introduction to Congestion Theory in Telephone Systems*, Edinburgh: Olner and Boyd, 1960.